Supplementary Material: The Edge of Depth: Explicit Constraints between Segmentation and Depth

Shengjie Zhu, Garrick Brazil, Xiaoming Liu
Michigan State University, East Lansing MI
{zhusheng, brazilga, liuxm}@msu.edu

1. Proof of Local Optimality

We give a brief proof that, under constructed transformation set $\{\phi(\mathbf{x} \mid \mathbf{q}, \mathbf{p})\}$, the proposed edge-edge consistency $l_c(\boldsymbol{\Gamma}(\mathbf{T_s} \mid \mathbf{T_d}),\ \mathbf{T}^*_{\mathbf{d}})$, can achieve the local optimality when the segmentation-augmented (or morphed) disparity edge points satisfy $\mathbf{T}^*_{\mathbf{d}} = \{\mathbf{p} \mid \left\|\frac{\partial \mathbf{I}^*_{\hat{\mathbf{d}}}(\mathbf{p})}{\partial \mathbf{x}}\right\| > \frac{t}{1+t} \cdot k_1\}$.

To prove this, let's start by evaluating the gradient of morphed disparity map $\mathbf{I}^*_{\hat{\mathbf{d}}}$ at a semantic edge pixel $\mathbf{q}$:

$$\begin{aligned}\forall \mathbf{q} \in \boldsymbol{\Gamma}(\mathbf{T_s} \mid \mathbf{T_d}),\ \left.\frac{\partial \mathbf{I}^*_{\hat{\mathbf{d}}}(\mathbf{x})}{\partial \mathbf{x}}\right|_{\mathbf{x}=\mathbf{q}} &= \frac{\partial \mathbf{I}_{\hat{\mathbf{d}}}(\phi(\mathbf{x}))}{\partial \phi(\mathbf{x})} * \left.\frac{\partial \phi(\mathbf{x})}{\partial \mathbf{x}}\right|_{\mathbf{x}=\mathbf{q}} \\ &= \left.\frac{\partial \mathbf{I}_{\hat{\mathbf{d}}}(\mathbf{y})}{\partial \mathbf{y}}\right|_{\mathbf{y}=\mathbf{p}} * \left.\frac{\partial \phi(\mathbf{x})}{\partial \mathbf{x}}\right|_{\mathbf{x}=\mathbf{q}},\end{aligned} \tag{1}$$

Note if $\mathbf{x} = \mathbf{q}$, $\phi(\mathbf{x} \mid \mathbf{q}, \mathbf{p}) = \mathbf{p}$. If $\left.\frac{\partial \mathbf{I}^*_{\hat{\mathbf{d}}}(\mathbf{x})}{\partial \mathbf{x}}\right|_{\mathbf{x}=\mathbf{q}}$ is sufficiently larger than a threshold, a semantic edge pixel $\mathbf{q}$ is also an edge pixel in the morphed disparity map, leading to the perfect edge-edge consistency for $\mathbf{q}$. We now derive the two terms in Eq. 1, in order to find that threshold.

When $\mathbf{x}$ is on the line segment $\overrightarrow{\mathbf{qp}}$, its projection $\mathbf{x}'$ overlaps with itself. We can thus compute $\left.\frac{\partial \phi(\mathbf{x})}{\partial \mathbf{x}}\right|_{\mathbf{x}=\mathbf{q}}$ as:

$$\begin{aligned}\left.\frac{\partial \phi(\mathbf{x})}{\partial \mathbf{x}}\right|_{\mathbf{x}=\mathbf{q}} &= \left.\frac{\partial \left(\mathbf{x} + \overrightarrow{\mathbf{qp}} - \frac{1}{1+t} \cdot \overrightarrow{\mathbf{qx}'}\right)}{\partial \mathbf{x}}\right|_{\mathbf{x}=\mathbf{q}} \\ &= \left.\frac{\partial \left(\mathbf{x} + \overrightarrow{\mathbf{qp}} - \frac{1}{1+t} \cdot \overrightarrow{\mathbf{qx}}\right)}{\partial \mathbf{x}}\right|_{\mathbf{x}=\mathbf{q}} \\ &= \left.\frac{\partial \left(\mathbf{x} + (\mathbf{p}-\mathbf{q}) - \frac{1}{1+t} \cdot (\mathbf{x}-\mathbf{q})\right)}{\partial \mathbf{x}}\right|_{\mathbf{x}=\mathbf{q}} \\ &= \left.\frac{\partial \left(\frac{t}{1+t} \cdot \mathbf{x} + \mathbf{p} - \frac{t}{1+t} \cdot \mathbf{q}\right)}{\partial \mathbf{x}}\right|_{\mathbf{x}=\mathbf{q}} \\ &= \frac{t}{1+t}.\end{aligned} \tag{2}$$

Using $\left.\frac{\partial \phi(\mathbf{x})}{\partial \mathbf{x}}\right|_{\mathbf{x}=\mathbf{q}} = \frac{t}{1+t}$ with Eq. 1, we have:

$$\begin{aligned}\forall \mathbf{q} \in \boldsymbol{\Gamma}(\mathbf{T_s} \mid \mathbf{T_d}),\ \left.\frac{\partial \mathbf{I}^*_{\hat{\mathbf{d}}}(\mathbf{x})}{\partial \mathbf{x}}\right|_{\mathbf{x}=\mathbf{q}} &= \left.\frac{\partial \mathbf{I}_{\hat{\mathbf{d}}}(\mathbf{y})}{\partial \mathbf{y}}\right|_{\mathbf{y}=\mathbf{p}} * \left.\frac{\partial \phi(\mathbf{x})}{\partial \mathbf{x}}\right|_{\mathbf{x}=\mathbf{q}} \\ &= \frac{t}{1+t} * \left.\frac{\partial \mathbf{I}_{\hat{\mathbf{d}}}(\mathbf{y})}{\partial \mathbf{y}}\right|_{\mathbf{y}=\mathbf{p}} \\ &> \frac{t}{1+t} \cdot k_1,\end{aligned} \tag{3}$$

where the inequality is derived from Eq. 1 of the main paper, which defines the threshold k_1 for detecting edge pixels on the original disparity map. Here, in morphed disparity map $\mathbf{I}^*_{\hat{\mathbf{d}}}$, since every counted semantic edge pixel $\mathbf{q} \in \boldsymbol{\Gamma}(\mathbf{T_s} \mid \mathbf{T_d})$ in computing the consistency l_c has a gradient magnitude larger than the threshold $\frac{t}{1+t} \cdot k_1$, $\mathbf{q}$ overlaps with the paired or matched depth/disparity edge pixel $\mathbf{p}$ as well, *i.e.*, $\mathbf{T}^*_{\mathbf{d}} = \{\mathbf{p} \mid \left\|\frac{\partial \mathbf{I}^*_{\hat{\mathbf{d}}}(\mathbf{p})}{\partial \mathbf{x}}\right\| > \frac{t}{1+t} \cdot k_1\}$. Thus, in morphed disparity map $\mathbf{I}^*_{\hat{\mathbf{d}}}$, semantic border overlaps with depth borders, making proposed consistency measurement l_c hit local minimum 0:

$$\begin{aligned}&\forall \mathbf{q} \in \boldsymbol{\Gamma}(\mathbf{T_s} \mid \mathbf{T_d}),\ \left.\frac{\partial \mathbf{I}^*_{\hat{\mathbf{d}}}(\mathbf{x})}{\partial \mathbf{x}}\right|_{\mathbf{x}=\mathbf{q}} > \frac{t}{1+t} \cdot k_1 \\ \Longleftrightarrow\ &\forall \mathbf{q} \in \boldsymbol{\Gamma}(\mathbf{T_s} \mid \mathbf{T_d}),\quad \delta(\mathbf{q}, \mathbf{T}^*_{\mathbf{d}}) = \min_{\{\mathbf{p} \in \mathbf{T}^*_{\mathbf{d}}\}} \|\mathbf{p} - \mathbf{q}\| \\ &\qquad\qquad\qquad\qquad\qquad\qquad\quad = \|\mathbf{q} - \mathbf{q}\| = 0 \\ \Longleftrightarrow\ &l_c(\boldsymbol{\Gamma}(\mathbf{T_s} \mid \mathbf{T_d}), \mathbf{I}^*_{\mathbf{d}}) = 0.\end{aligned} \tag{4}$$

This shows that, under the defined transformation, we are realigning the depth edge set $\boldsymbol{\Omega}$ to the segmentation edge set $\boldsymbol{\Gamma}^{\mathbf{d}}_{\mathbf{s}}$, making the edge-edge consistency a local optimality.

Note that the threshold $\frac{t}{1+t} \cdot k_1$ is not actually being applied to the morphed disparity map for edge detection. Rather, we derive it as the condition that will be naturally satisfied in our work, when both the morph function and k_1 threshold for disparity map depth estimation (Eq. 1 of the main paper) are employed.

Depth Decoder						
layer	**k**	**s**	**c**	**res**	**input**	**activation**
upconv5	3	1	256	32	econv5	ELU[1]
iconv5	3	1	256	16	$\uparrow$ upconv5, econv4	ELU
upconv4	3	1	128	16	iconv5	ELU
iconv4	3	1	128	8	$\uparrow$ upconv4, econv3	ELU
disp4	3	1	1	1	iconv4	Sigmoid
upconv3	3	1	64	8	iconv4	ELU
iconv3	3	1	64	4	$\uparrow$ upconv3, econv2	ELU
disp3	3	1	1	1	iconv3	Sigmoid
upconv2	3	1	32	4	iconv3	ELU
iconv2	3	1	32	2	$\uparrow$ upconv2, econv1	ELU
disp2	3	1	1	1	iconv2	Sigmoid
upconv1	3	1	16	2	iconv2	ELU
iconv1	3	1	16	1	$\uparrow$ upconv1	ELU
disp1	3	1	1	1	iconv1	Sigmoid

Table 1: The network architecture of our decoder. **k**, **s** and **c** denote the kernel size, stride and output channel numbers of the layer, respectively. **res** refers to relative downsampling scale to the input image. $\uparrow$ symbol means a $2\times$ nearest-neighbour upsampling to input.

2. Network details

Across our experiments, we use ImageNet [2] pretrained ResNet18 and ResNet50 [6] as our encoder. Our decoder structure is same as Godard *et al.* [5] and Waston *et al.* [10], as detailed in Table 1. We also incorporate other practices such as color augmentation, random flip, edge-aware smoothness and exclusion of stationary pixels.

3. More Ablations

In this section, we perform additional ablations to further validate our proposed approach. We ablate (1) Our proposed morph strategy achieves local optimality of edge-edge consistency l_c, and (2) The stereo occlusion mask $\mathbf{M}$ boosts clear borders. All our ablations are conducted on Eigen [3] test splits of KITTI [4].

Reducing edge-edge consistency via morphing: We plot the edge-edge consistency loss l_c under various edge detection thresholds k_1 in Fig. 1. We cross-validate morphing (detailed in main paper Section **3.1**) as a technique to achieve local optimality of l_c from Fig. 1 via showing consistently decreased measurement l_c after applying morphing once and twice. The lower loss in Fig. 1 shows that our models are more consistent with segmentation compared to [10]. Additionally, increased threshold k_1 leads to thinner edges and neglects distant objects, which have two effects. First of all, thinner edges make edge-edge consistency to be more challenge, thus higher loss values. Second, focusing on close-range objects can best leverage the high-quality segmentation, which leads to larger improvement margin over the baseline [10].

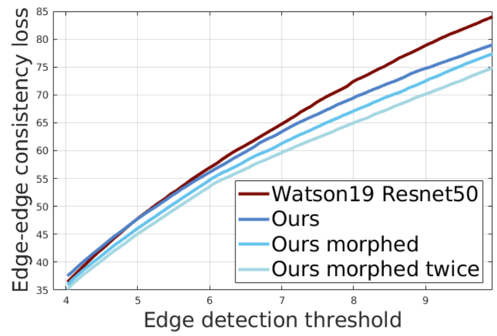

Figure 1: We plot the edge-edge consistency l_c between Watson19 [10] and ours at different edge detection thresholds k_1. Additionally, we show the change of consistency l_c after applying morph strategy once and twice during inference, in addition to using our learned network.

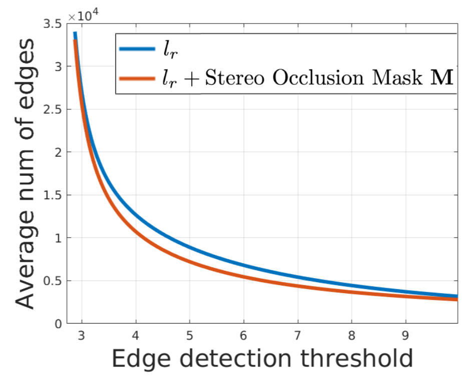

Figure 2: The effects of proposed stereo occlusion mask $\mathbf{M}$. We plot the trend of the average detected edge numbers $\frac{1}{n}\sum_{i=1}^{i=n}(\|\frac{\partial \mathbf{I}_{\mathbf{d}}^{\mathbf{i}}(\mathbf{x})}{\partial \mathbf{x}}\| > k_1)$ at different edge detection thresholds k_1, where n is for total number of tested images.

Stereo Occlusion Mask: In Fig. 5, we observe bleeding artifacts universally exist in stereo-based systems [8, 9, 10]. In [10], the utilization of stereo proxy label partially suppresses it as its additional constrain on the low texture area. [5] reduces the artifacts via supervision from videos. In comparison, without any additional supervision sources, we eliminate it via the proposed stereo occlusion mask $\mathbf{M}$. As an example, the top-right subfigure of Fig. 3 reveals a clearer and thinner border when comparing l_r against $l_r + \mathbf{M}$. This motivates us to treat "thinness" as a measurement and use the average detected edge number $\frac{1}{n}\sum_{i=1}^{i=n}(\|\frac{\partial \mathbf{I}_{\mathbf{d}}^{\mathbf{i}}(\mathbf{x})}{\partial \mathbf{x}}\| > k_1)$ as an approximated metric of border clearance, as shown in Fig. 2. As expected, after applying the mask $\mathbf{M}$, edges become more "thinner" and clearer, reflected as the decreased number of detected edges.

More quality comparisons: We show additional qualitative examples when different loss are applied in Fig. 3. We further provide qualitative comparisons against the baseline method [10] in Fig. 4, and other methods in Fig. 5.

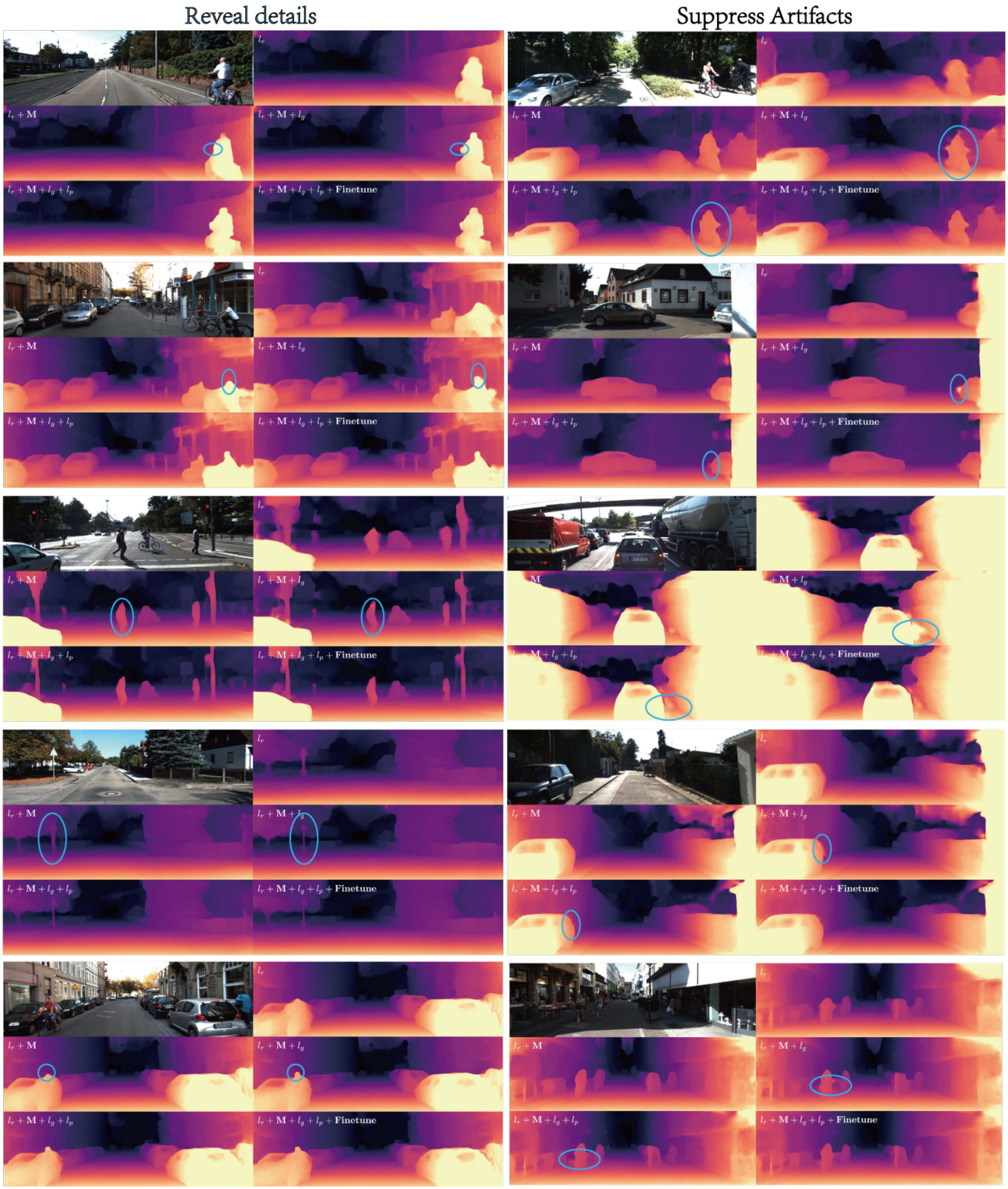

Figure 3: On the left column, explicit utilization of segmentation information helps recovering more details. On the the right, we show blobbed border artifacts in the low texture areas, caused by noisy predicted segmentation labels and low constrain from the photometric loss l_r. We suppress the artifacts by the incorporation of texture weight $\mathbf{w}$ and utilization of proxy stereo labels [7, 10].

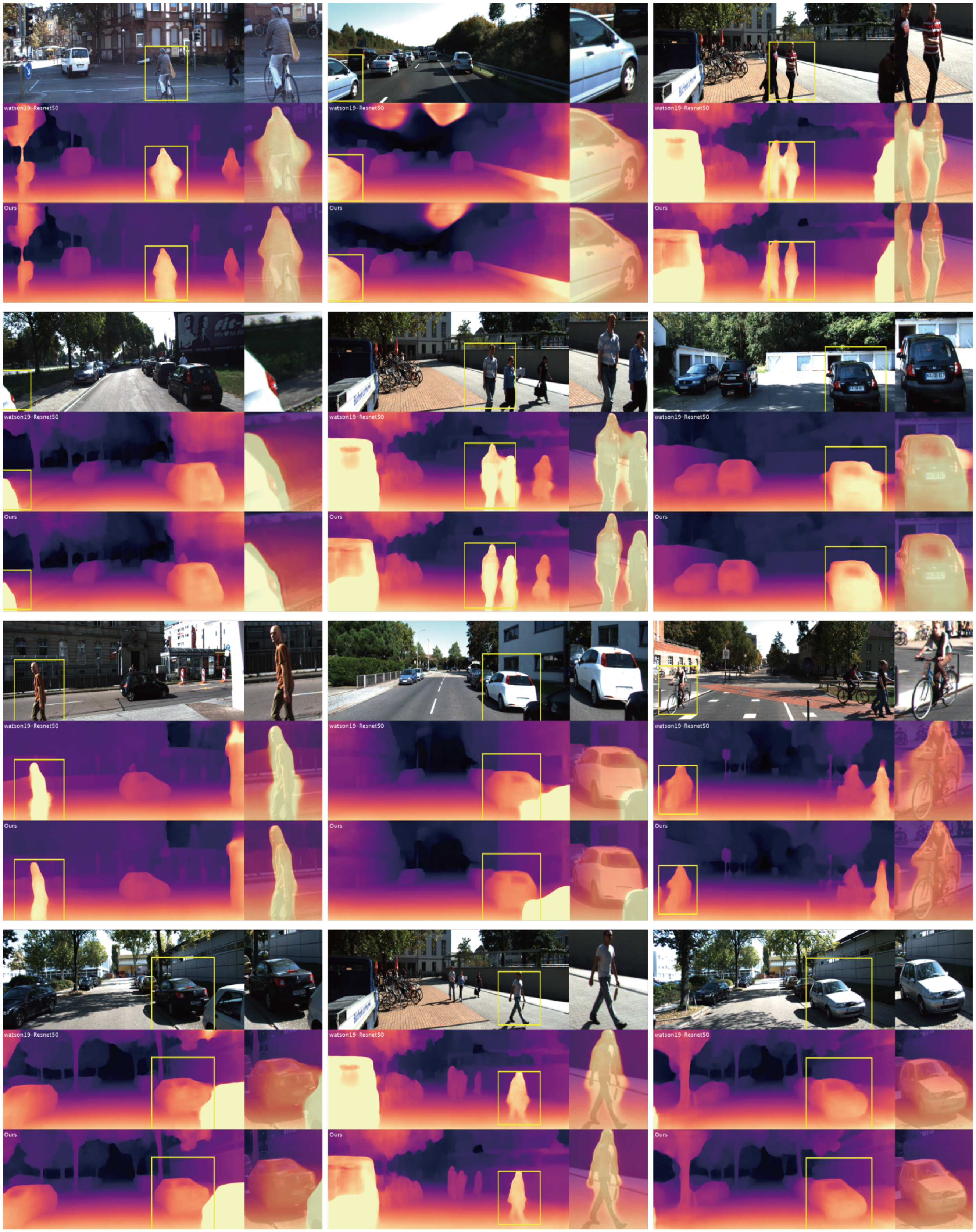

Figure 4: More comparison between ours model and the state-of-the-art baseline [10]. Content within yellow box is zoomed in and attached to the right. We show significantly improved border quality compared to the method of [10].

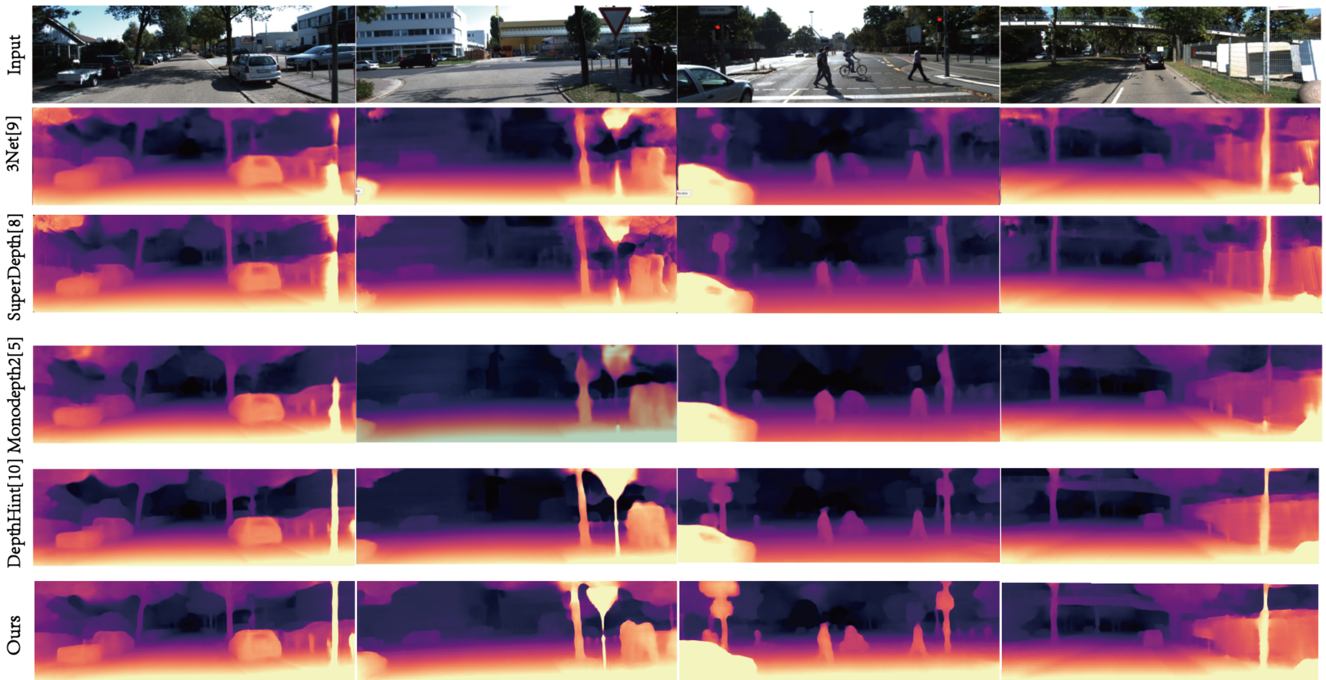

Figure 5: Comparison against other state of the arts [5, 8, 9, 10]. Our method reconstructs more object details compared to previous works and possesses the most clear border overall.

References

[1] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

[3] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems (NIPS)*, pages 2366–2374, 2014.

[4] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361, 2012.

[5] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3828–3838, 2019.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[7] Yevhen Kuznietsov, Jorg Stuckler, and Bastian Leibe. Semi-supervised deep learning for monocular depth map prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6647–6655, 2017.

[8] Sudeep Pillai, Rareş Ambruş, and Adrien Gaidon. Superdepth: Self-supervised, super-resolved monocular depth estimation. In *Proceedings of the International Conference on Robotics and Automation (ICRA)*, pages 9250–9256, 2019.

[9] Matteo Poggi, Fabio Tosi, and Stefano Mattoccia. Learning monocular depth estimation with unsupervised trinocular assumptions. In *Proceedings of the IEEE International Conference on 3D Vision (3DV)*, pages 324–333, 2018.

[10] Jamie Watson, Michael Firman, Gabriel J Brostow, and Daniyar Turmukhambetov. Self-supervised monocular depth hints. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2162–2171, 2019.